Astronomical Data Analysis Software and Systems XXI
ASP Conference Series, Vol. 461
Pascal Ballester, Daniel Egret, and Nuria P. F. Lorente, eds.

Long Term Preservation of Data Analysis Software at the NASA/IPAC Infrared Science Archive

Harry I. Teplitz,[1] Steven Groom,[1] Timothy Brooke,[1] Vandana Desai,[1] Diane Engler,[1] John Fowler,[1] John Good,[1] Iffat Khan,[1] Deborah Levine,[1] and Anastasia Alexov[2]

[1]*Infrared Processing and Analysis Center, Caltech, Pasadena, CA 91105, USA*
[2]*Astronomical Institute Anton Pannekoek, The Netherlands*

Abstract. The NASA/IPAC Infrared Science Archive (IRSA) curates both data and analysis tools from NASA's infrared missions. As part of our primary goal, we provide long term access to mission-specific software from projects such as IRAS and Spitzer. We will review the efforts by IRSA (and within the greater IPAC before that) to keep the IRAS and Spitzer software tools current and available. Data analysis tools are a vital part of the Spitzer Heritage Archive. The IRAS tools HIRES and SCANPI have been in continual use since the 1980's. Scanpi offers a factor of 2 to 5 gain in sensitivity over the IRAS Point Source Catalog by performing 1D scan averaging of raw survey data at specified arbitrary position. In 2007 SCANPI was completely modernized, with major code revisions. HIRES returns IRAS survey images with higher resolution than the IRAS Sky Survey Atlas (ISSA). We are currently undertaking a modest revision to the tool to ensure continued reliability. In the next two years, the US Planck Data Center plans to adapt both tools for use with Planck data, and deliver them to IRSA for long term curation.

1. Motivation

Data analysis tools are a major part of archival research. Expert users frequently require tools to extend existing data sets in a variety of ways. Not all scientific needs can be met with a one-size-fits-all processing. In addition, data tools enable the use of new methods and calibration as processing techniques evolve. Thus, it is the responsibility of data archives to ensure the availability of data analysis tools in the long term.

This long-term curation is a challenge. Computer technology continues to evolve. Cutting-edge data tools in one decade may need significant upgrades to remain relevant over time. Advances in processing techniques also motivate a demand for new functionality. To complicate the situation, the original developers of tools often move on to other projects and eventually retire. Archives must then find another way to retain the necessary institutional memory.

In this paper, we describe the multi-decade effort to curate a suite of data tools at the Infrared Processing Analysis Center (IPAC) and the NASA/IPAC Infrared Science Archive (IRSA). IRSA curates and serves infrared scientific data products from NASA's infrared and submillimeter projects and missions. It is the archive for the Infrared Astronomical Satellite (IRAS; Neugebauer et al. (1984)) and 2 Micron All-Sky Survey (2MASS; Skrutskie et al. (2006)) datasets. More recently, IRSA has expanded to be-

come the permanent home of the Spitzer Heritage Archive (SHA), the WISE Archive, and the NASA Planck Archive. IRSA will provide seamless access to Herschel data, and long-term access to public data from SOFIA. As part of our charter, IPAC and IRSA have curated the data tools needed to support analysis of the archival holdings.

In our discussion of tool curation, we will start with the analysis tools from IRAS. Since their initial development in the 1980's, these tools have required maintenance and upgrades as technology has evolved. Recently, IRSA has assumed responsibility for another large set of data analysis tools – those within the Heritage Archive for the Spitzer Space Telescope (Werner et al. 2004). Finally, we will discuss plans to adapt these same tools for use with future projects.

2. IRAS Data Analysis Tools

The Infrared Astronomical Satellite (IRAS) was a joint project of the US, UK and the Netherlands. The IRAS mission performed an unbiased, sensitive all sky survey at 12, 25, 60 and 100 microns. In an 11 month mission in 1983, IRAS surveyed 96% of the sky. IRAS detected about 350,000 infrared sources.

The IRAS focal plane consisted of apertures leading to integrating cavities (see Figure 1). The detector masks were rectangular in aspect and infrared sources scanned across the focal plane parallel to the narrow dimension of the detectors in all observational modes. As the satellite scanned, astronomical objects crossed the vertical "array" of sensors. The measurements were reported as Time Ordered Data. With the known speed of the scan, shapes can be reconstructed.

A complication in the analysis of IRAS data was the varying response of individual detectors as objects scan across them. The response for each detector can be calibrated but must be taken into account during processing. Figure 2 shows the cross-scan profiles of the 25 micron detectors.

There are three key points to keep in mind when considering IRAS data: (1) The data are not images, and were never intended to make images. IRAS was designed to reliably detect point sources by scanning across them with individual detector elements; (2) the scan geometry varied greatly depending on target location; and (3) The response of each individual detector depended on a complicated sensitivity function, the illumination history, and the scan geometry.

Two data analysis tools were developed for users of IRAS data. The first, "SCANPI", offers a factor of 2 to 5 gain in sensitivity over the IRAS Point Source Catalog by performing 1D scan averaging of raw survey data at specified arbitrary position. The second, "HIRES", returns IRAS survey images with higher resolution than the IRAS Sky Survey Atlas (ISSA).

2.1. SCANPI

SCANPI, the Scan Processing and Analysis tool, is an interactive software tool for viewing, plotting and averaging the calibrated survey scans from the Infrared Astronomical Satellite (IRAS); these scans are the fundamental data from the IRAS survey. SCANPI is useful for measuring the fluxes of extended, confused or faint sources, for diagnosing source extent, and for estimating local upper limits. The sensitivity gain, which is comparable to that obtained in the IRAS Faint Source Survey, is a factor of 2-5 over the IRAS Point Source Catalog (PSC), depending on the local noise and number

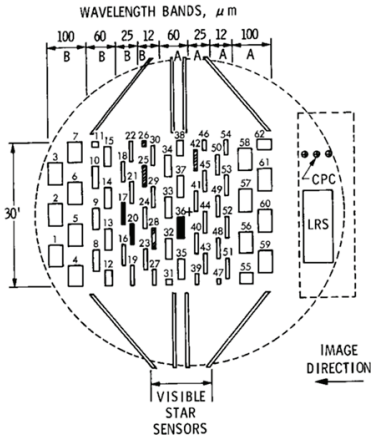

Figure 1. A schematic drawing of the IRAS focal plane. The numbered rectangles in the central portion each represent an aperture above a filter and field lens leading into an integrating cavity containing a detector. The image of a source that crossed the focal plane in the Y direction as indicated. The filled-in detectors were inoperative while the cross-hatched detectors showed degraded performance during the mission. Taken from the IRAS Explanatory Supplement, Fig II.C.6 (Beichman et al. 1988)

of scans crossing the target position. SCANPI allows users to interactively subset, plot and coadd IRAS scan data at any sky position. Source lists can also be uploaded.

2.2. HIRES

HIRES uses the Maximum Correlation Method (MCM; Aumann et al. (1990)) to produce images with better than the nominal resolution. It is a powerful tool for studying morphology and separating confused sources. HIRES produces IRAS resolution better than an arcminute (5x increase over baseline) with fluxes good to 20% (due to background determination uncertainties). Output mosaics are available as 1x1 or 2x2 degree images, with 15 or 30 arcsecond pixels. The user controls which scans are used as input, and what de-striping (background and time-varying detector response correction) technique is used. Figure 3 shows the output of HIRES processing of the IRAS observation of M31.

MCM is based on comparing simulated images to actual measurements. The idea is that of a simulator that could use a trial image and an observational model to predict what the measurements would be, compare those to actual measurements, and use the differences to feed back a correction to the trial image. This process is then iterated until it converges. MCM handles the case that every measurement (sensor or pixel) has a different point response function (PRF), as in the case of IRAS (Figure 2). It makes use of noise measurements in the comparison. MCM operates on overlapping measurements with, if necessary, completely under-sampled PSFs. It reduces to Richardson-Lucy algorithm for isoplanatic oversampled PSF with no prior noise estimates

The need to take individual PRFs into account is well summarized by the following anecdote from co-author J. Fowler an early implementation of MCM:

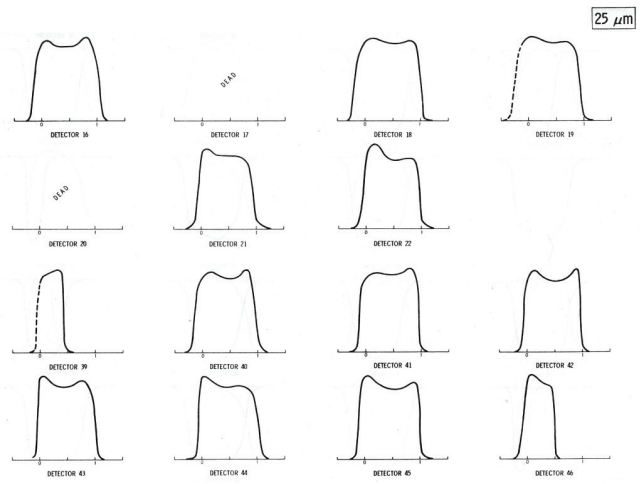

Figure 2. Cross-scan profiles of each 25 micron detector deduced from scans across NGC6543. The measurements were made in the pointed mode where cross scan position is well known. Each scan has been normalized to unity at its peak value. Taken from the IRAS Explanatory Supplement, Fig IV.A.3.1 (Beichman et al. 1988)

"When Mike Melnyk and I were ready to try executing it for the very first time in full-capacity mode, we used a raster scan of the asteroid Egeria which had been serving as one of the photometric standards. We did not yet have these realistic response functions, so we were using "top-hat" (i.e., rectangular) response functions corresponding to each detector's aperture size. We ran a 60-micron case first, expecting a crash with lots of error messages, but the run finished normally and produced surprisingly good images of Egeria. The images got sharper as more iterations were done, and we were amazed at how well it all worked right off the bat. After we finished congratulating ourselves, we decided to try the 25-micron channel. To our dismay, the images came out with two Egerias! It didn't take long to figure out what had happened: the real response functions (all but one) have a dip in the middle; the only way MCM could explain the observations using rectangular response functions was to have two sources of light. Once we got the real response functions into the program, we got back to a single Egeria. The lack of a dip in the 60-micron response functions had allowed a single Egeria with top-hats in that channel."

2.3. The 1980's

SCANPI grew out of the IRAS data analysis pipeline. The mission processing included the software package ADDSCAN, which combined multiple scans of a given position to increase sensitivity. The need for individual investigators to define the input scans became apparent as the breadth of IRAS archival research increased (e.g. Helou et al. (1988)). Thus SCANPI was developed and made part of the final delivery of the IRAS archive.

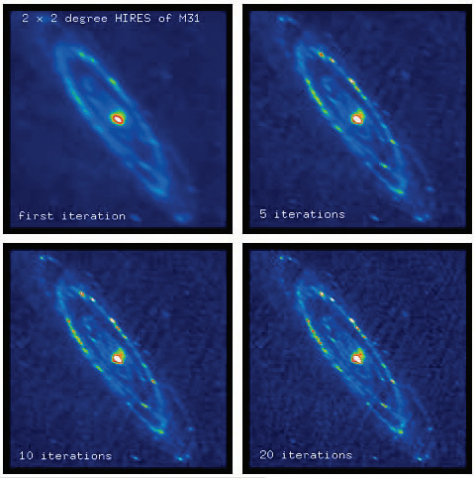

Figure 3. A 2×2 degree IRAS 100 micron image of M31, processed by HIRES with increasing numbers of iterations.

SCANPI originally ran on the Cyber computer at IPAC. Initially its output was provided to users as a paper rather than digital product.

HIRES was also developed for the same hardware. Initially, running the algorithm was interactive, given the need to deal with options for "destriping" (removal of time varying signatures from the sky background and/or detector). HIRES was run by an operator at IPAC, who would take the input parameters specified by the user. The output needed to be carefully examined to determine whether additional iteration was necessary. This quality analysis (QA) step was initially done using contour plots of the resulting intensity map. Once the operator was satisfied with the QA results, the output maps would be sent to the user.

2.4. The 1990's

SCANPI and HIRES were ported to the Sun/Unix platform in 1991 and 1992, respectively. When the porting was complete, SCANPI output began to be supplied in digital form rather than on paper. At this time, careful benchmarking was performed to ensure that the change in architecture did not impact the results. The storage medium for the output, including the benchmark results, was 8mm tapes.

At this time, HIRES was upgraded to allow the HIRES operator to perform QA by examining images rather than simply contour maps. This improvement is an early example of the need to keep data tools current with evolving visualization capabilities. Later, circa 1995, limited resources precluded devoting time for extensive QA of each HIRES run by the operator.

The 1990's also saw significant improvement in HIRES functionality. The FITS header of output images was upgrade to include the –CAR specification for projection type. The evolving standardization of FITS provides another example of the need for keeping imaging tools current. Consider also the move from the CDELT specification to the use of the CD matrix.

In 1996, the HIRES destriping algorithm was upgraded. More effective methods of removing electronic baseline and destriping were developed (e.g., the IRAS Galactic

Plane images; Cao et al. (1997)). The nature of the data prevented complete automation, with the user needing to examine the output to check. As noted above, resources did not permit as much QA effort from the operator.

At the end of the decade, with the inception of IRSA, curation of HIRES and SCANPI at IPAC were given a new home.

2.5. 2001-2011

This decade saw a resurgence of interest in IRAS, to support research with Spitzer. There was sufficient demand for new scan processing that the Spitzer project funded a major upgrade of SCANPI. The modernization was completed in 2007.

SCANPI was given an new, web-based user interface to facilitate iterative processing. Figure 4 shows an example of the new SCANPI. In addition, a program-friendly interface was added to allow non-interactive calls. Program friendly interfaces have long been a standard at IRSA, and are becoming ever more important in the age of the Virtual Observatory.

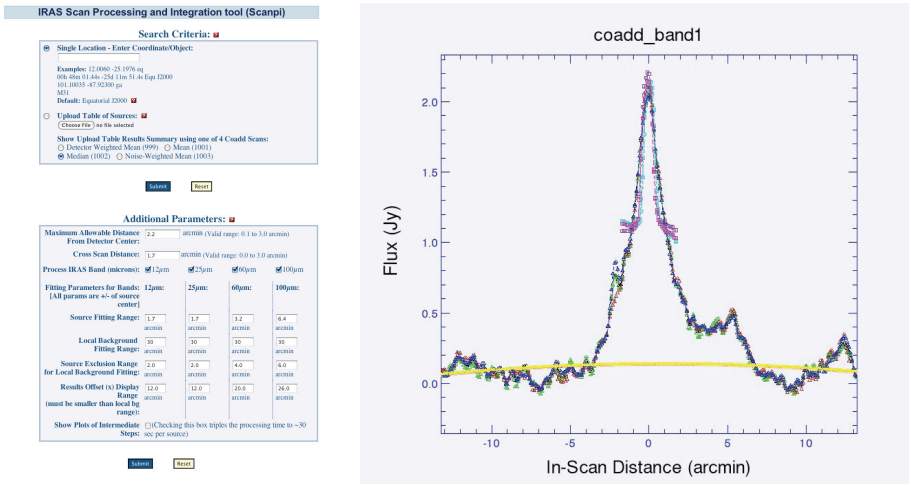

Figure 4. (left) The input web-page for the SCANPI tool at IRSA; (right) The output of SCANPI processing of the IRAS scans of the core of M31.

In addition to the interfaces, the SCANPI upgrade included a significant restructuring of the code. Decades of small fixes by multiple developers had led to unnecessarily complicated code. In addition bug fixes and other small changes had sometimes been performed without sufficient documentation. Several developers had moved on to other projects or other institutions, leading to a loss of institutional memory. Having a single developer take control of SCANPI for the upgrade provided a much needed clean-up, and greatly increased the longevity of the tool.

HIRES continued to be in use during the decade. In 2007, there were 5,000 separate requests for HIRES images. Many of these were to support the use of IRAS maps in combination with GALEX mosaics. Boissier et al. (2007) used this combination to test models of extinction in large, local galaxies. HIRES operation at IPAC continued to be labor intensive during these peak usage periods, but resources did not permit a full overhaul of the code.

The MCM algorithm (see Section 2.2), which is the basis of HIRES, was also applied to other applications. In particular, it became the basis of the coaddition pipeline for the Wide-field Infrared Survey Explorer (WISE; Wright et al. (2010)). Figure 5 shows an example of the output of the WISE imaging pipeline.

In just the past two years, IRSA has begun the process of updating the HIRES code for use on newer platforms. This was motivated in part by a hardware crash in 2008 that took HIRES offline briefly. The binaries of the HIRES code ran perfectly well on the circa-1997 hardware that had been their home. At first try, they did not run on a modern Sun computer. Fortunately, several other computers of the same vintage were soon identified, and the problem was solved temporarily.

Now, as a first step, the code was upgraded so that it will compile on a modern Sun computer. In addition it is being ported to run on a Linux platform as well. All of IRSA's other software infrastructure has migrated from Sun to Linux in recent years. Finally, HIRES is being converted to remove the need for an operator. User-specified parameters will be passed directly to the modules. Users will be more responsible for correcting any mistakes that are caught by the input validation module. Output QA will continue to be the users responsibility. As part of the process, the original regression testing benchmarks (still stored on 8mm tape) are being used to ensure that no functionality is lost. The tapes are still readable, but reading them was an important reminder that proper storage of benchmark results is necessary. Tapes will degrade over time.

Figure 5. The WISE 3-color Atlas image of NGC1514.

2.6. The Future

IRSA is planning for continued use of the HIRES and SCANPI tools. The skeptical reader may ask why we continue to perpetuate these tools rather than attempting to bring more modern tools to bear. First, the IRAS data are quite particular, and these tools are optimized for that data set. And, just as importantly, there simply aren't resources to design or adapt something new – astronomical archives operate on limited budgets, and we necessarily use the tools that are given to us by the missions that take the data.

In addition to analysis of IRAS data, these tools will now be adapted for use with data from the Planck mission (Planck collaboration et al. 2011). Planck is similar to IRAS, with nearly identical survey strategy and a focal plane filled with heterogeneous, not-diffraction-limited detectors. The US Planck Data Center at IPAC, together with IRSA, will leverage our IRAS experience by adapting HIRES and SCANPI to Planck data. SCANPI will be used to extract all scans through a given celestial position; the scans will be individually available, baseline-removed, and coadded. HIRES will use all scans through a rectangular patch of sky to construct an image with enhanced resolution in a finely-gridded mosaic whose effective beam will be much less variable over the image than what results from simple co-addition using the asymmetric Planck response functions. The new versions of the tools will be closely tied to IRSA infrastructure and user interface (UI).

In addition, the WISE coaddition pipeline will be adapted to become a user tool, the WISE Custom Coadder. This will enable coaddition based on epoch. A simple web interface will be developed, in the context of existing IRSA UI, including HIRES underlying middleware to manage data exchange with the Level "1b" image archive that is available as part of other IRSA services.

3. Spitzer Data Analysis Tools

Given its diverse nature, Spitzer data cannot be optimally processed with an automated procedure in all cases. The Spitzer Science Center (SSC) pipelines removed most instrumental signatures, but many users require additional processing by advanced software tools. Support for the complete set of tools developed for Spitzer is outside the scope of IRSA's budget, so the tools are divided into Tier I (absolutely required) and Tier II ("best effort" support). See the Spitzer documentation website[1] for a full description of the instruments and tools.

TIER 1

- **MOPEX/APEX:** Mosaicking and source extraction for all Spitzer imaging (IRAC, MIPS, IRS-PeakUp). Enables IRAC photometry at the few percent level, accounting for intra-pixel sensitivity. Standard software (e.g., daophot) is less effective.
- **SPICE:** IRS staring-mode spectral extraction; simultaneous visualization of slit-position in 2D dispersed image and a direct image.
- **CUBISM:** IRS mapping-mode spectral extraction; construction of data cubes; CUBISM supplies a unique functionality for which there is no accessible alternative.
- **GeRT:** MIPS 70/160 pipeline allowing for custom re-reduction; reduction is particularly sensitive to observing details.
- **CUPID:** IRS pipeline; vital option for power users.
- **IRSCLEAN:** Cleans bad pixels from 2D IRS spectra.

[1] `http://irsa.ipac.caltech.edu/data/SPITZER/docs/`

TIER 2

- **BANDMERGE:** Merges catalogs, taking into account positional uncertainties.
- **DARK SETTLE:** Mitigates time-dependent dark current in IRS spectra.
- **IRS FRINGE:** De-fringes IRS spectra.
- **PAHFIT:** Fits PAH lines in IRS spectra.

3.1. Documentation

Documentation of data tools, for both users and developers, has been a vital part of the process of transitioning Spitzer tools from the operating mission to the long-term archive at IRSA. A detailed plan was developed with SSC. This plan included: User Manuals; Tutorials for each of the tools, with detailed science examples for a variety of use cases – the "data analysis cookbook".

In addition, developer documentation has been essential. In the early years of MOPEX work, responsibility for the code passed to a new developer when the original author left the project. The resulting learning curve motivated an increased attention on documentation. For each analysis tool, SSC will provide a developer overview document, in addition to careful comments in c-code to ease modification of the source code. Another set of crucial documents is build instructions to explain how to make the binaries. Finally, for each of the tools a Regression Testing Plan was developed. These plans included data sets and name lists (parameter definition files) that fully exercise the code, along with outputs and scripts to find discrepancies with previous versions.

4. Science User Support

The NRC report "Portals to the Universe" (2007) strongly recommended that NASA should maintain mission expertise at the archive centers for the long-term support of both novice and power users. A key aspect of support for the SHA will be the retention of instrument expertise to support archival research.

In the past, IRSA has had the luxury of relying on the "institutional memory" at IPAC and a Great Observatory on-site. In the new downsizing environment, we will lose that expertise if it is not funded via IRSA. Experience has shown that the highest demand for instrument expertise is during the 5 years following the end of mission, after which the effort can be reduced.

5. Conclusions and Lessons Learned

Long-term preservation of data analysis tools are a crucial part of the astronomical archive mission. IPAC and IRSA experience have been responsible for the software tools from the IRAS mission, and more recently the Spitzer Space Telescope. We have discussed the important points in the history of these tools, and a number of lessons have been learned.

First and foremost, retention of institutional memory is required to maintain software tools. This includes both knowledge of the code and science expertise to support users. Key factors in accomplishing this include:

- Documentation of the use of the tools (User Manuals, Tutorials)
- Code documentation, including developer guides and build instructions
- Benchmarking software output to enable future regression testing, including proper storage of benchmark results (tapes can become unreadable/degrade).

As software tools are maintained, archives must necessarily make trade-offs. Resources are often scarce and limit what upgrades are possible. So choices must be made whether a required upgrade is worth it, or should the tool finally be retired. IRSA has not made that choice yet in the case of IRAS and Spitzer tools, but it is a concern for the future. This choice will involve evaluation of the cost of the upgrade, the demand for the tool, and the relevance of the data set to be analyzed.

Data analysis tools will be easier to maintain, and have longer life, if they can take advantage of modern archival infrastructure. There are significant advantages to generalizing a given tool beyond use with a single mission data set, such as the expansion of HIRES to WISE and Planck.

Finally, it is hard to predict the improvements that will occur over decades in computer technology. We have seen that major computer changes occur on a much faster time scale than the turn over in relevant data sets. It is vital that tools be adapted to take advantage of new capabilities.

Acknowledgments. The tools discussed in this paper are the result of the talent and dedication of dozens of scientists and software engineers. We are happy to acknowledge the hard work of the teams that strive to ensure the best possible tools are available for analysis of NASA's infrared data.

This research has made use of the NASA/ IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

References

Aumann, H. H., Fowler, J. W., & Melnyk, M. 1990, AJ, 99, 1674
Beichman, C. A., Neugebauer, G., Habing, H. J., Clegg, P. E., & Chester, T. J. 1988, Infrared astronomical satellite (IRAS) catalogs and atlases. Volume 1: Explanatory supplement 1
Boissier, S., et al. 2007, ApJS, 173, 524
Cao, Y., Terebey, S., Prince, T. A., & Beichman, C. A. 1997, ApJS, 111, 387
Helou, G., Khan, I. R., Malek, L., & Boehmer, L. 1988, ApJS, 68, 151
Neugebauer, G., et al. 1984, ApJL, 278, 1
Planck collaboration, et al. 2011, A&A, 536, A1. 1101.2022
Skrutskie, M. F., et al. 2006, AJ, 131, 1163
Werner, M. G., et al. 2004, ApJS, 154, 1
Wright, E. L., et al. 2010, AJ, 140, 1868